UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

MA-INF 4222 - Natural Language Processing - Lab

Chaitali Prabhu (3069208)

Debanjali Biswas (3063994)

Hemanth Kumar Reddy Mayaluru (3063488)

Kaushikee Khaund (3068762)

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

QA over Knowledge Graphs:

Overview:

1. Entity Detection & Linking for questions (Hemanth Kumar Reddy Mayaluru)
2. Answer generation for a given question (Chaitali Prabhu)
3. Semantic coherence calculation (Kaushikee Khaund)
4. Recommended Questions generation (Debanjali Biswas)

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

1. Entity Detection & Linking

- Data extraction & preprocessing
  - Dataset: **DBpedia**
  - Using **SPARQL** queries, entity labels and types are extracted from DBpedia
  - Data preprocessing to make it available in the format needed to train the model
- Entity Detection from a given Question
  - New model is built using **spacy**
  - Trained the model using the training dataset
  - Saved the model and tested with the test dataset
- Entity Linking
  - Data preprocessing
  - Using the **AskNow** API entity linking is done on the dataset

Training Model

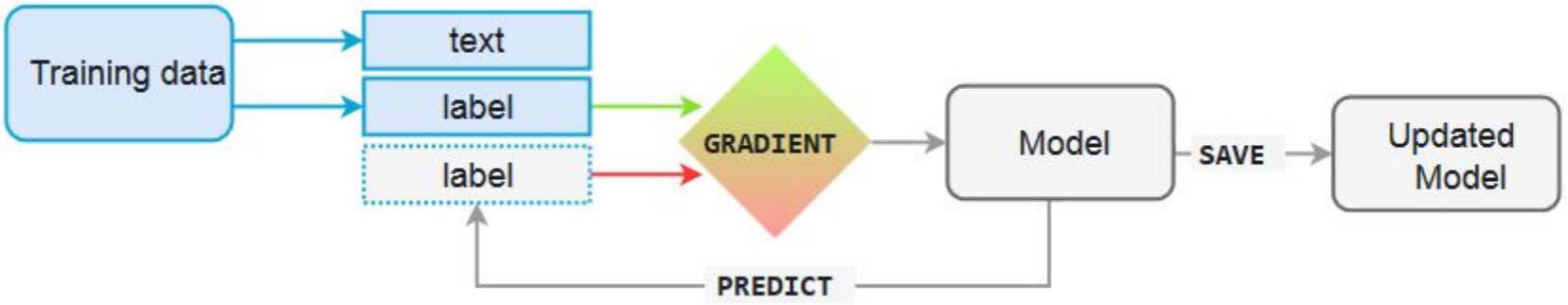


Figure 1: retrieved from: *https://spacy.io/usage/training*

SAMPLE_TRAIN_DATA =
[('which genre of album is harder faster ', {'entities': [(24, 31, 'Album')]}), ('what movie is produced by Warner Bros ', {'entities': [(26, 37, 'Company')]}),]

Model Processing

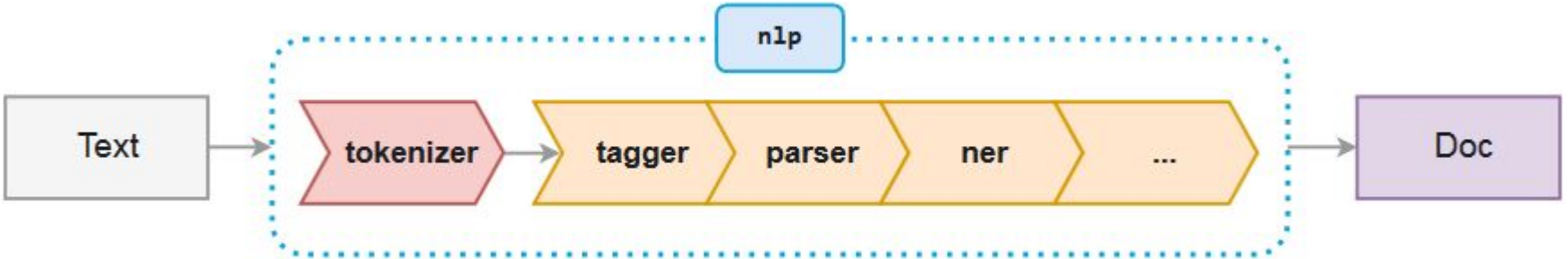


Figure 2: retrieved from: *https://spacy.io/usage/*processing-pipelines

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Example

Question:

‘what movie is produced by Warner Bros’

Model Output:

Entities:[('Warner Bros', 'Company')]

Tokens[('what', '', 2), ('movie', '', 2), ('was', '', 2), ('produced', '', 2), ('by', '', 2), ('Warner', 'Company', 3), ('Bros', 'Company', 1)]

Entity Linking Output:

Link Types : ['relation', 'entity', 'relation']

‘Word : Type’ = [{'movie': 'relation', 'Warner Bros': 'entity', 'produced': 'relation'}]

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Dataset & Tech stack:

- Dataset: DBpedia
- Python 3
- SPARQL
- Spacy
- EARL

Issues Faced

- Model Training took too much time
- Fix: mini-batches

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Results:

Dataset	Precision	Recall	f-score	Precision (Tokens)
Train	99.274	99.330	99.302	100.0
Test	84.467	83.041	83.748	100.0

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \qquad recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

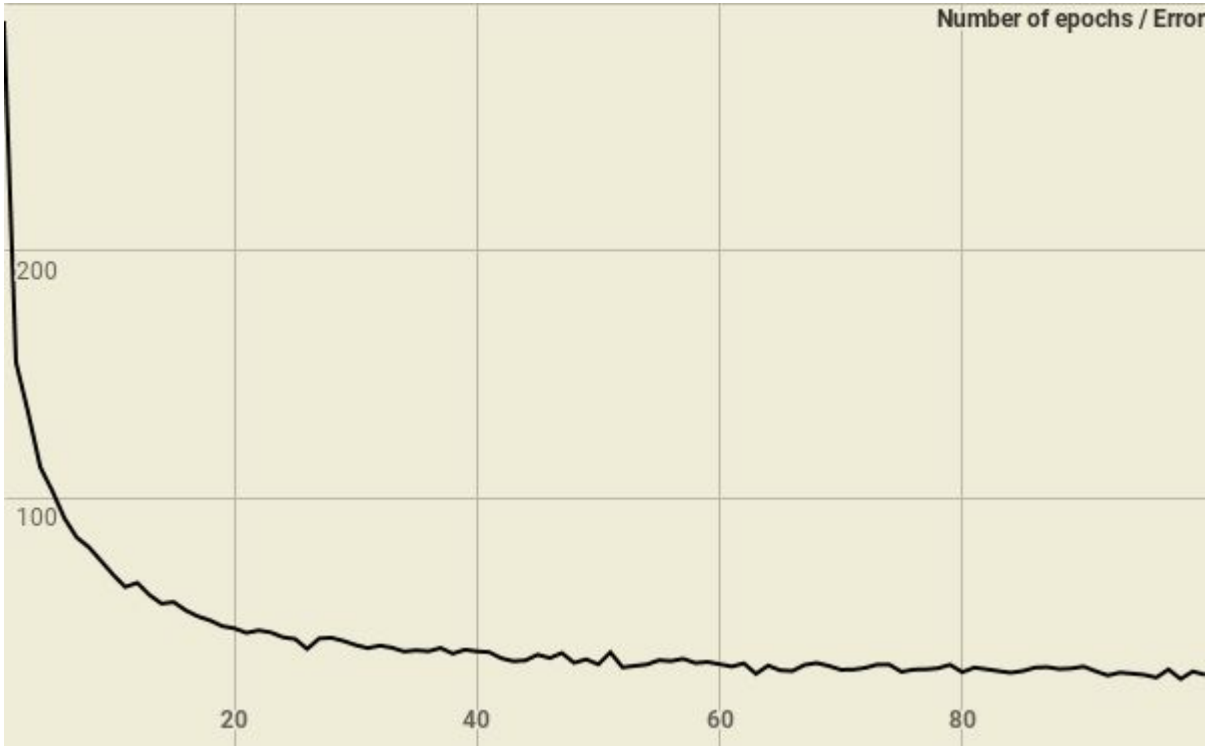
Results - error vs #epochs

1st iter:

292.7562237191361

100th iter:

28.838385919308397



UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

References:

- Dubey, Mohnish & Banerjee, Debayan & Chaudhuri, Debanjan & Lehmann, Jens. (2018). EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs.
- https://spacy.io/usage/processing-pipelines
- Salman Mohammed, Peng Shi, and Jimmy Lin. Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 291-296, May 2018, New Orleans, Louisiana

Future Work:

- Train the model for still larger datasets
- User friendly GUI can be developed

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

2. Answer generation for a given question

To retrieve answers for simple questions from DBPedia and translate them to sentences.

- Get direction of question
- POS Tagging on question to get predicate and subject/object
- SPARQL to get answer
- Translate triple to sentence.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Dataset & Tech stack:

- Dataset: DBpedia
- Python 3
- SPARQL
- NLTK
- Sklearn

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Methodology

- Get direction of question
  - Using logistic regression to predict the direction of question.
  - Training dataset: https://raw.githubusercontent.com/castorini/SimpleDBpediaQA/master/V1/train.json
- POS Tagging on question to get predicate and subject
  - Using nltk to get the verbs and nouns from the question.
  - Verbs and common nouns : potential predicates
  - Proper nouns : subject
  - Using edit distance to get closest predicate from the predicates of subject.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Methodology

- Use SPARQL to get answer for given Subject and Predicate
  - Querying dbpedia to get answer.
- Forming sentences for answers
  - Using basic sentence template to generate answers.
  - Get noun form of predicate.
  - The <predicate> of <subject> is <object>.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Result : for 1 answer

```
Direction =  ['forward']
Question   what is the official language of United States ?
POS  [('what', 'WP'), ('is', 'VBZ'), ('the', 'DT'), ('official', 'JJ'), ('language', 'NN'), ('of', 'IN'), ('United', 'NN
P'), ('States', 'NNPS'), ('?', '.')]
Noun =  United States
Verb  is
Other  official_language
http://dbpedia.org/resource/United_States
{'Wikipage page ID': 'http://dbpedia.org/ontology/wikiPageID', 'Wikipage revision ID': 'http://dbpedia.org/ontology/wiki
PageRevisionID', 'Link from a Wikipage to another Wikipage': 'http://dbpedia.org/ontology/wikiPageWikiLink', 'Link from
a Wikipage to an external page': 'http://dbpedia.org/ontology/wikiPageExternalLink', 'thumbnail': 'http://dbpedia.org/on
tology/thumbnail', 'area total (km2)': 'http://dbpedia.org/ontology/PopulatedPlace/areaTotal', 'population density (/sqk
m)': 'http://dbpedia.org/ontology/populationDensity', 'has abstract': 'http://dbpedia.org/ontology/abstract', 'anthem':
'http://dbpedia.org/ontology/anthem', 'area total (m2)': 'http://dbpedia.org/ontology/areaTotal', 'capital': 'http://dbp
edia.org/ontology/capital', 'demonym': 'http://dbpedia.org/ontology/demonym', 'ethnic group': 'http://dbpedia.org/ontolo
gy/ethnicGroup', 'flag (image)': 'http://dbpedia.org/ontology/flag', 'founding date': 'http://dbpedia.org/ontology/found
ingDate', 'government type': 'http://dbpedia.org/ontology/governmentType', 'language': 'http://dbpedia.org/ontology/lang
uage', 'largest city': 'http://dbpedia.org/ontology/largestCity', 'leader': 'http://dbpedia.org/ontology/leader', 'leade
r title': 'http://dbpedia.org/ontology/leaderTitle', 'longName': 'http://dbpedia.org/ontology/longName', 'motto': 'http:
//dbpedia.org/ontology/motto', 'official language': 'http://dbpedia.org/ontology/officialLanguage', 'percentage of area
water': 'http://dbpedia.org/ontology/percentageOfAreaWater', 'population total': 'http://dbpedia.org/ontology/population
Total'}
pred  official language
http://dbpedia.org/ontology/officialLanguage
tokens  JJ
Answer =  The  official language  of  United States is  Federal government of the United States
```

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Result : for multiple answer

```
Direction =  ['forward']
Question   who is the creator of Batman ?
POS  [('who', 'WP'), ('is', 'VBZ'), ('the', 'DT'), ('creator', 'NN'), ('of', 'IN'), ('Batman', 'NNP'), ('?', '.')]
Noun =  Batman
Verb  is
Other  creator
http://dbpedia.org/resource/Batman
{'birth place': 'http://dbpedia.org/ontology/birthPlace', 'Wikipage page ID': 'http://dbpedia.org/ontology/wikiPageID',
'Wikipage revision ID': 'http://dbpedia.org/ontology/wikiPageRevisionID', 'Link from a Wikipage to another Wikipage': 'h
ttp://dbpedia.org/ontology/wikiPageWikiLink', 'Link from a Wikipage to an external page': 'http://dbpedia.org/ontology/w
ikiPageExternalLink', 'has abstract': 'http://dbpedia.org/ontology/abstract', 'creator (agent)': 'http://dbpedia.org/ont
ology/creator'}
pred  creator (agent)
http://dbpedia.org/ontology/creator
tokens  NN
Length  2
['agents']
Answer =  The  creator agents  of  Batman are  Bob Kane, Bill Finger
```

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Issues

- Generating answers for backward questions
  - Issue in retrieving the backward predicate.
  - Can be solved with relevant dataset of predicates with backward predicates.
- Multiple values for answers
  - If the object has multiple values, it considers it as different values.
- Entity alias
  - If the entity has alias names, redirection to valid uri does not work.
- Sentence generation
  - Uses a fixed template.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Future Work

- Finding closest predicate from dbpedia
  - Using edit distances, now using syntactic similarity.
  - Use cosine similarity to compare natural language phrases and predicates in dbpedia to find semantically closer predicates.

Accuracy

For logistic regression : 94.4970%

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

3. Semantic Coherence Calculation

- Finding view counts of the Entities
  - Dataset: **DBpedia**
  - **SPARQL** query is used to extract labels from DBpedia
  - Getting the view counts of the Entities from Index files given.
  - Comparison of the view counts for each Entity.
- Getting all the Predicates
  - Dataset: **DBpedia**
  - **SPARQL** query is used to extract labels of predicates from DBpedia
  - Forward and Backward predicates labels for that Entity

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Semantic Coherence Calculation

- Finding Semantic Similarity
  - Vector forms of the predicate labels using **SPACY**
  - Then the **Cosine Similarity Function** is used to find the similarity.
  - The predicate having lowest cosine distance is the most similar.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Example

Data provided from the previous part

Entity1 = http://dbpedia.org/resource/Germany

Entity2 = http://dbpedia.org/page/Berlin

Predicate(label) = capital

Output for Entity:

maximum index is for entity1: http://dbpedia.org/resource/Germany

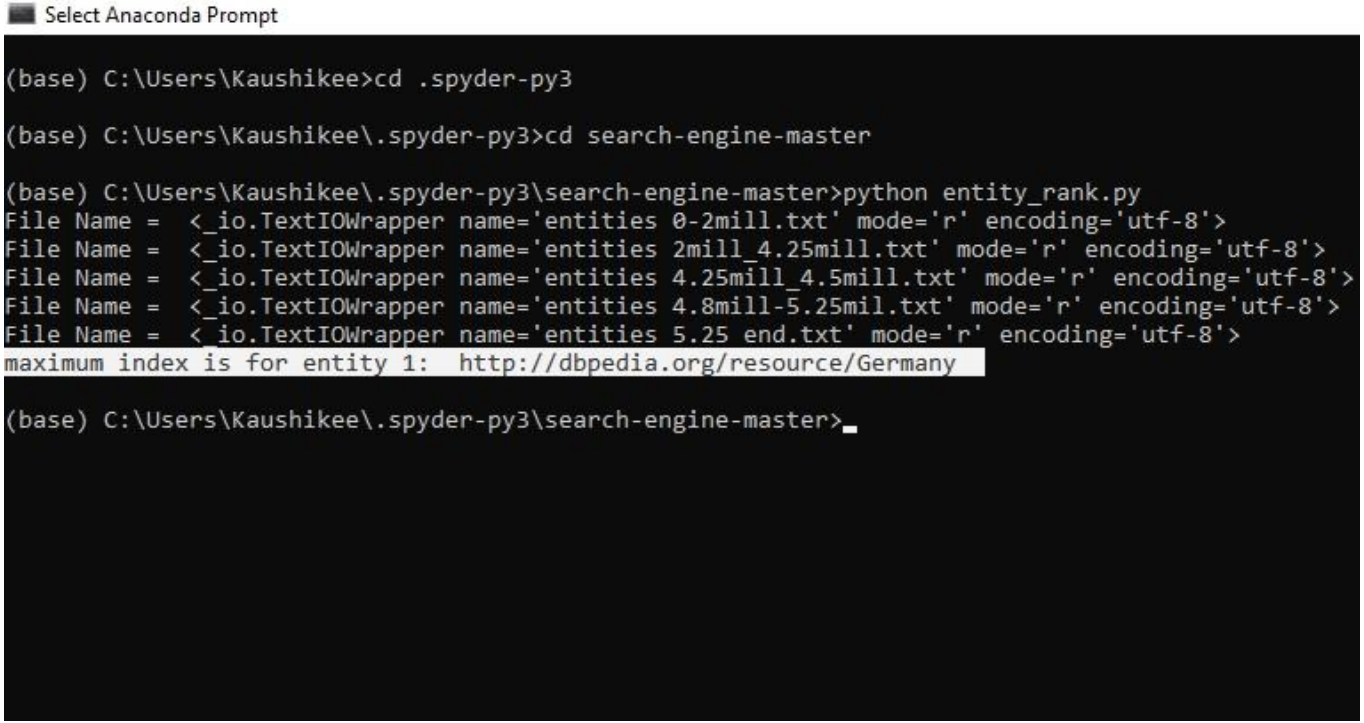
Output for Cosine Similarity of Predicate Labels:

minimum value is for backward predicate: 0.21539670432115066

resultant backward predicate: city

Result

Result for Entity with highest view count



Result

Result for Predicate which is most similar to the given predicate

```
minimum value is for backward predicate:  0.21539670432115066
index for backward predicate:  18
resultant backward predicate:  city

(base) C:\Users\Kaushikee\.spyder-py3\search-engine-master>
```

```
Backward Pred:  type
Backward Pred:  victim (resource)
Backward Pred:  wine region
forward  [0.43056495 0.43056495 0.45070434 0.45070434 0.64600916 0.64600916
 0.58665798 0.58665798 0.3289992  0.3289992  0.46425029 0.46425029
 0.45744099 0.45744099 0.71796295 0.71796295 0.39394    0.39394
 0.5066923  0.5066923  0.         0.         0.34493595 0.34493595
 0.45716316 0.45716316 0.50890067 0.50890067 0.37854647 0.37854647
 0.3380784  0.3380784  0.23288224 0.23288224 0.36957538 0.36957538
 0.3038477  0.3038477  0.34443963 0.34443963 0.42634317 0.42634317
 0.31931206 0.31931206 0.45744099 0.45744099 0.30718111 0.30718111
 0.38825602 0.39770579 0.48906372 0.47347815 0.5297176  0.34545803
 0.48673117 0.29208599 0.793897   0.49598062 0.46274814 0.44809999
 0.72561324 0.37908478 0.34745742 0.36887081 0.43627617 0.35473193
 0.51138415 0.43137545 0.42914089 0.56460764 0.42467816 0.57475081
 0.54379012 0.49602526 0.72295874 0.61720595 0.73324003 0.5442207
 0.34608954 0.46676416 0.43674465 0.61552873 0.41035218 0.7295546
 0.36979214 0.54907865 0.49793752 0.43211918 0.50896324 0.39783322
 0.44750136 0.36473195 0.33268122 0.34065029 0.47054267 0.32049333
 0.41901751 0.44511623 0.49571932 0.37228848]
backward  [0.35840971 0.42135408 0.64600916 0.51445842 0.45871641 0.38006594
 0.39620953 0.37364891 0.29438293 0.27987729 0.45758394 0.71483788
 0.40013342 0.40068384 0.3699951  0.22241414 0.49054035 0.26121918
 0.2153967  0.25270386 0.31201216 0.297294   0.28769849 0.32369488
 0.25956359 0.41709032 0.39723082 0.42382343 0.4859193  0.39939258
 0.35837729 0.44028617 0.32846767 0.44553168 0.43538038 0.32922797
 0.37780823 0.69950643 0.31035137 0.74106608 0.23288224 0.47256234
 0.25701009 0.30033496 0.48841824 0.40989546 0.44071389 0.41704108
 0.32437128 0.25502973 0.25450986 0.39966398 0.30379437 0.30027453
 0.32416605 0.29304514 0.25151204 0.35793365 0.35495933 0.38324312
 0.39957088 0.29548242 0.80625    0.3238148  0.53701098 0.43330812
 0.31017551 0.42206547 0.37224914 0.35138485 0.31524866 0.34988736
```

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

DataSet and Technology Used:

- DBpedia
- Python 3
- SPARQL
- SPACY
- Cosine Similarity Function

Future work:

- Any efficient similarity function other than Cosine Similarity Function it can be used.
- The comparison only works for two entities.

4. Recommended Questions generation

- **Data preprocessing**
  - Dataset: **RevisedDBpediaQA (extension of the SimpleDBpediaQA)**
  - Dataset contains **Triple or Facts (Subject, Object, Predicate)** , **Questions** and **Direction (Forward or Backward)**
  - Data preprocessing is done using scripts (from the Dataset) to convert it in the format needed to train the model
  - Creating Input Vocabulary using **Gensim**
- **Task Definition**
  - Transfering **Facts into Questions**
- **Model**
  - **Seq2Seq Model** for Question Generation from Triples

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Seq2Seq Model

- Encoder
  - Fact Embedding
- Decoder
  - GRU Recurrent Neural Network with Attention Mechanism
- Model is based on the methodology explained in the paper:

  “Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus”

Model Architecture

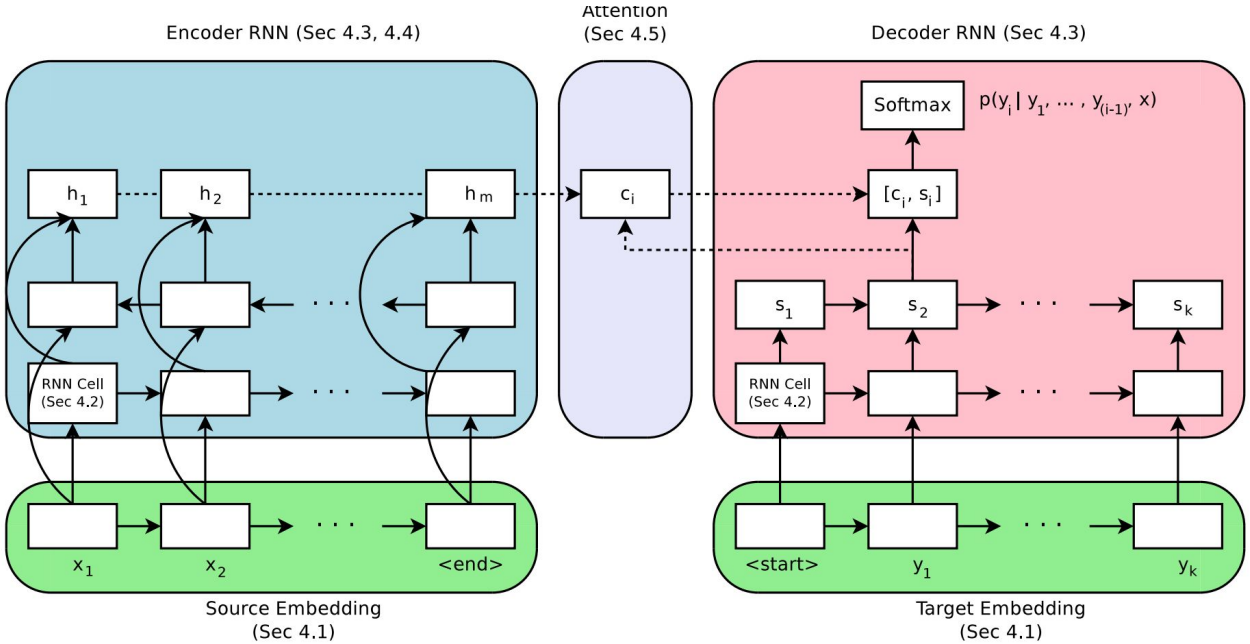


Figure 1: Taken from “Massive Exploration of Neural Machine Translation Architecture”.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Example :

- Input:
  - Fact :
    - Subject: http://dbpedia.org/resource/Germany
    - Predicate: http://dbpedia.org/ontology/capital
    - Object: http://dbpedia.org/resource/Berlin
  - Direction:
    - Forward

- Output: “what is the capital of germany”

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Tech Stack

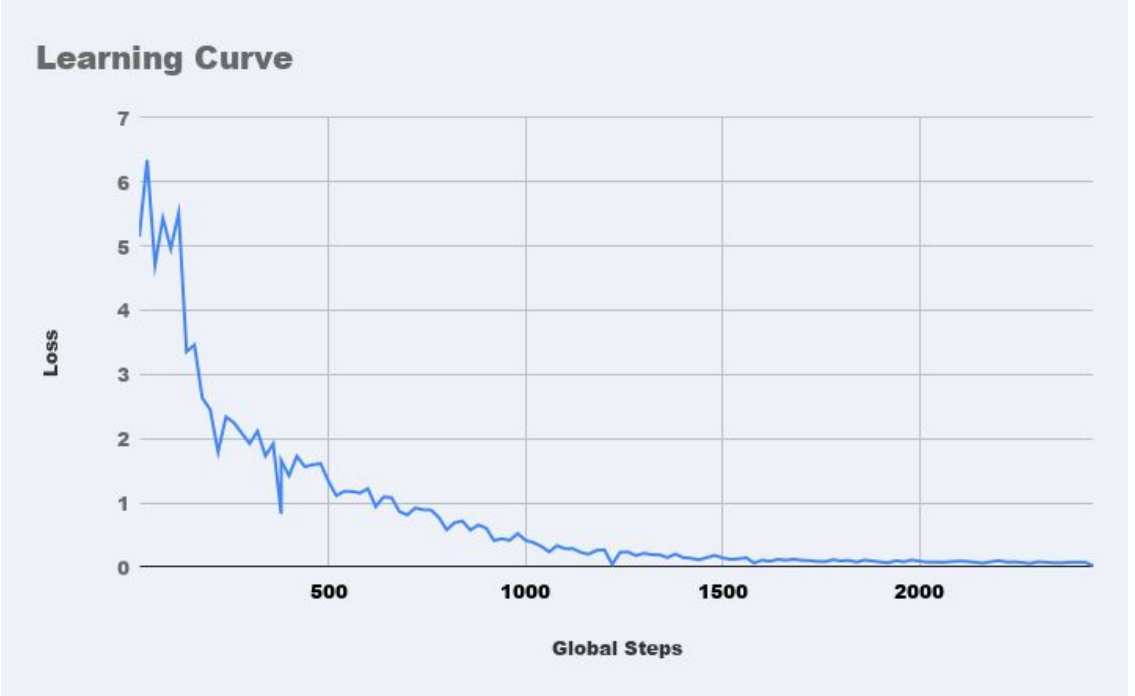
- Python 3
- Tensorflow
- Gensim

Issues Faced

- Dataset Preprocessing: Unwanted lines of Data in the Dataset
  - Fix: Manually removing unwanted lines from the Dataset
- Model doesn’t work for out-of-vocabulary words and unseen Predicates

Training

- Total no. of lines in the training dataset: 30,077 (after removal)
- No. of iterations : 2
- Each iteration :
  - No. of Epochs: 20
  - Size of Batches: 500
- Loss after training: 0.01723878



UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Results

- Input:
  - Fact :
    - Subject: http://dbpedia.org/resource/Rock_music
    - Predicate: http://dbpedia.org/ontology/genre
    - Object: http://dbpedia.org/resource/Anniversary_(Split_Enz_album)
  - Direction:
    - Backward
- Output: “album is a type of rock music”

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

References

- Serban, I., García-Durán, A., Gülçehre, Ç., Ahn, S., Chandar, A.P., Courville, A.C., & Bengio, Y. (2016). Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus. *CoRR, abs/1603.06807*.

- ElSahar, H., Gravier, C., & Laforest, F. (2018). Zero-Shot Question Generation from Knowledge Graphs for Unseen Predicates and Entity Types. *NAACL-HLT*.

- Britz, D., Goldie, A., Luong, M., & Le, Q.V. (2017). Massive Exploration of Neural Machine Translation Architectures. *CoRR, abs/1703.03906*.

UNIVERSITÄT BONN

SMART DATA ANALYTICS
FROM DATA TO KNOWLEDGE

Thank You !